Determining Author or Reader: A Statistical Analysis of Textual Features in Children's and Adult Literature

Lindsey Geybels

University of Antwerp, Prinsstraat 13, 2000 Antwerpen, Belgium

Abstract

Due to the nature of literary texts as being composed of words rather than numbers, they are not an obvious choice to serve as data for statistical analyses. However, with the help of computational techniques, words can be converted to numerical data and certain parts of a text can be examined on a large scale. Textual elements such as sentence length, word length and lexical diversity, which are associated by scholars on the one hand with the writing style of an individual author and on the other with the complexity of a text and the intended age of its readers, can thus be subjected to statistical evaluation. In this paper, data from little under 700 English and Dutch books written for different ages is analysed using a statistical linear mixed model. The results show that the textual elements studied are better qualified to detect the age of the intended reader of a text than the identity or age of the author.

Keywords

children's literature, linear mixed models, authorship attribution, readership attribution, text complexity

1. Introduction

Statistical tests are traditionally associated with exact science or sociology; for example to ascertain the effect of a treatment on the growth of a certain plant or to determine the efficacy of a school-based smoking prevention curriculum. Over the past few decades, the applicability of statistics in making quantitative decisions has expanded through innovations in technological, and specifically computational, areas. While its subject matter contains a myriad of data that can be used as input for computational analysis, literature is one of the fields not traditionally studied through mathematical calculations. However, the use of computational techniques allows literary researchers to "change slippery words into more absolute numerical [...] substitutes" [30]. Possibly the most popular implementation of this method is found in the field of stylometry, which occupies itself with the study of linguistic style. By applying statistical analysis to specific features of a set of texts, stylometry is often employed to attribute authorship, either to anonymous or disputed documents. Due to the nature of this research, which often studies a handful of texts by as many authors to determine their stylistic proximity to an anonymous text, the analyses tend to be small-scaled [24] with a focus on authorship attribution of general literature.

This paper fits into the CAFYR (Constructing Age For Young Readers) research project and

CHR 2022: Computational Humanities Research Conference, December 12 – 14, 2022, Antwerp, Belgium

lindsey.geybels@uantwerpen.be (L. Geybels)

0000-0002-6557-924X (L. Geybels)

borrows from it the extensive corpus of 692 titles of children's, young adult, and adult literature written by Dutch, English and Flemish authors. To avoid the traditional treatment of the study of children's literature as being isolated from literature for adults, the corpus is focused on crosswriters; authors who write for children of different ages as well as an adult readership. Both the focus on these authors and the number of texts exceed the scope of many previous studies. For each title in the corpus, the average sentence length, average word length, measure for lexical diversity and ratio of dialogue versus narration is extracted and used as input for a linear mixed model. Linguistic features have been used in stylometric analysis to study several pieces of metadata. First and foremost, while the analysis of word frequencies, most often in the form of function words or n-grams, is generally considered to be more reliable in the practice of authorship attribution [3, 14], previous research has proposed the use of simple token-based lexical features and vocabulary richness as markers to quantify an author's writing style [22, 10, 4]. Second, questions of chronology within the oeuvre of one author, corresponding to the changes in the writing style of an author as they age, have been answered by so-called stylochronometry [18]. Finally, the same features are also often used to determine the readability of a text, a practice that is more common in children's literature and is connected to the age of the intended readership. In a stylometric analysis of the oeuvres of ten British, Dutch and Flemish authors, Haverals et al. found that the style of texts often correlate with the age of the intended reader [15]. To judge if the textual features listed above, used both for authorship attribution and text readability, are better at predicting the author, their age or that of the reader, this paper reports on statistical analyses using linear mixed models while also investigating a possible difference in language. The question this paper tries to answer, in other words, is whether sentence length, word length, lexical diversity and the ratio between dialogue and narration are better suited to determine issues of authorship or of readability.

2. Previous research

Previous research using statistical techniques to analyse children's literature is scarce but not absent. Roger Clark identified a striking level of interest of feminist social science in a quantitative approach to children's literature [5]. His literature review contains a list with over thirty articles featuring research into sexism, sex-roles, gender and other feminist social topics by 'counting', as Clark puts it somewhat simplistically [16, 9]. While not all of these studies include statistical tests in their methodology, a number of them do and the field of social sciences seems to be in a leading position when it comes to statistical analyses of children's literature [8]. In general literature, quantitative analyses of textual features have mainly been employed to study authorial style. In the field of computational stylometry, authorship attribution has been a hot topic for several decades, with different researchers defending the validity of certain textual features, while discarding others, in the search of a text's true author. Because anonymous or pseudonymous authors are rarely found in children's literature, it has yet to be considered as fruitful data for quantitative studies of textual features. However, these analyses can reveal much more about a text than merely the author's subconscious writing style. Haverals et al. used a stylometric analysis on the oeuvres of ten authors who write English or Dutch books intended for different readerships, ranging from 'middle child' (children aged 6 to 8) to

'adult' (people aged 18 and up). They looked not only for clustering according to the individual author, but included the age of the author and the age of the intended reader as features to their analyses. From several case-based analyses as well as from the corpus-wide analysis they performed on English titles, they concluded that stylometric analysis can be fruitful when investigating issues of, what I will call, readership attribution; the task of identifying the age of the intended readership of a text [15].

Most often, stylometric analyses are based on frequencies of function words, as is the case for Haverals et al.'s study. However, in children's literature, textual features including sentence length, word length and lexical diversity are more commonly used to determine the writing style or complexity of a text. These features are often consciously manipulated by the author or editor of a book in connection with literacy and education. Guidelines for the readability of a text, or readability formulas such as ARI, Gunning fog and Dale-Chall in large part rely on these values to categorise texts according to an appropriate reading age. Scholarly interest in this topic has produced mainly small-scale studies which do not rely on statistical analyses [28, 11, 25]. A notable exception is Celia Catlett Anderson's dissertation on style in children's literature. She performs a statistical analysis of several textual features from both books written for children and for adults, including sentence and word length, lexical repetition and the amount of dialogue, corresponding to the features studied in this paper. This paper aims to expand on previous research by using a larger dataset of textual features from 692 texts. Furthermore, whereas most studies into writing style have focused on a single language, the analyses below look for any significant differences between Dutch and English texts, based on sentence length, word length, lexical diversity and the ratio between dialogue and narration.

The main idea underlying studies into authorship attribution in the field of stylometry is that "authors have an unconscious aspect to their writing style" [17], certain features that they include in their writing without being able to actively manipulate them. Even before the emergence of computational methods to aid in the quantitative study of texts, the analysis of simple token-based lexical features, including sentence length and word length, was being used to attribute authorship [27, 33]. Although the length of lexical units is generally not deemed to be a reliable indicator to determine a text's authorship, there have been studies that nuance this view [26]. Furthermore, the distribution of word length was one of the elements used by Patrick Juola in 2013, when he brought the study of authorship attribution to the attention of people outside of literature and linguist departments [21]. When it comes to the study of text complexity in children's literature and, closely connected to it, the age of the intended reader of a text, sentence length is a more commonly used measure. According to Colleen Lennon and Hal Burdick, "the best predictor of the difficulty of a sentence is its length" [23]. In a study of children's book publishers, Celia Catlett Anderson relates the average length of sentences in a book to the age of its readership. Through interviews with several publishers, she found one of the most common requests to children's books' authors is to shorten sentences to make the text simpler and more accessible to young readers [1]. Guidelines like these determine the reading level of books. However, as each child develops reading skills at its own pace, this does not translate directly to the age of the reader but there is a close connection between both. This suggests that the average sentence length of a text is often, contrary to the assumption of early studies in authorship attribution, under conscious control of the author.

Lexical diversity, as a measurement of the number of different words in a text, is most simply

represented by the ratio of the number of unique words (types) to the total number of words (tokens) [20]. Several more complex formulae exist to measure lexical diversity, which in addition consider the number of words that occur a specific number of times, their frequencies and arbitrary constants [14]. Parallel to the lexical features of sentence and word length, the vocabulary richness of a text has been disputed as well as confirmed as a reliable indicator in the search of a text's author [4, 13]. In addition to the standard authorship attribution problem, Holmes identified a second use for stylometric research: chronological problems, or the hypothesis that stylistic features develop during an author's life [17]. It was this that Tallentire addressed in a study based on the ratio between hapaxes and tokens to measure lexical diversity, where he concluded that lexical diversity decreases with age of the author [31]. Victoria Johansson confirms this correlation in her study conducted on the writing and speech of children aged 10, 13, 17 and adults. However, from a developmental point of view, she found that lexical diversity increases with age, attesting to vocabulary development in children [19]. Lexical diversity is not only linked to the creation of a text but also its consumption and the pedagogical role of children's literature. In Wanner et al.'s 'Age Suitability Analysis', several aspects that play a part in determining the appropriate reading age of a book are listed [32]. One of them is linguistic complexity; the "difficulty of the writing style" which is measured using a variation on type/token ratio. Texts intended for younger readers contain more repetitions and thus will present a more limited vocabulary [2].

The ratio between dialogue and narration is not traditionally connected to studies of authorship or text complexity, but according to Rita Ghesquière, this measure is influenced by the age of the intended reader, as she has found that books for younger readers usually contain more dialogue when compared to books aimed at adults [12].

3. Collecting data

The analyses in this paper are conducted on 692 works of prose fiction published by Dutch, British and Flemish authors between 1970 and 2020. To investigate the relation of metadata concerning author and reader to textual features, the following information is stored for each book: language, year of publication, the age of the author at the time of publication and the age of the intended reader. Due to a movement in the United Kingdom which opposes children's literature publishers' strategy of 'banding' their books by adding explicit age guidelines to the cover, the latter was found on the physical books for only a small percentage of the texts, which were almost exclusively Dutch books. For the remaining texts, the age of the intended reader was retrieved from other sources, if available, in order of precedence: publisher's catalogues, author's websites, the Dutch database for children's books (CBK) or websites of booksellers.

After collecting metadata on the extensive corpus, several features were extracted from the individual texts, which were first stripped of paratext and chapter headings: sentence length, word length, lexical diversity, and the ratio between direct and indirect speech. For the first two, tokenizers from the Natural Language Toolkit were used. To calculate lexical diversity, the Moving Average Type-Token Ratio (MATTR) was used [7]. By computing the average of the type/token ratio of a moving window with fixed length, this method resolves the fallacy of many other formulae which do not consider the length of the text being analysed. The text

was first lemmatized using spaCy, which is sensitive to different parts of speech. Covington recommends a small window size when analysing patterns of repetition, or a large window when trying to determine the size of the writer's vocabulary [6]. In the analyses for this article, the window size is set to 1000 words, a value close to the word count of the shortest book in the corpus, *Sinclair Wonder Bear* by Malorie Blackman (1036 words). The ratio between direct and indirect speech was added to the present study as a control feature as it has not been considered in authorship attribution studies, and thus no correlation with the author is assumed. The distinction between narration and dialogue in the texts in the corpus was made by means of quotation marks. Unfortunately, a large part of the texts was digitized from scans and small punctuation marks are subject to errors due to the OCR (Optical Character Recognition) process. To minimize this error as much as possible, manually annotated material was used as input where available. About one third of the corpus is annotated to study aspects of characterisation in the broader research project. These annotations include the identification of characters' speech versus narration and can thus be used to extract the number of words in either category.

4. Method: why mixed models

Basic statistical models, such as linear regression, are suitable to analyse simplistic and clean data, which is a rarity in life sciences. Much more often, data is non-independent, which arises from repeated measures or a hierarchical structure; a study can rely on multiple measurements of the same subjects at different times or data points can be grouped. Ignoring non-independence by using standard linear regression in these cases means that not all the variation in the data will be captured. Linear mixed models (LMM) provide an answer to this problem by considering random effects as well as the fixed effects recognised by linear statistical models and compiling all individual results into one model. LMM distinguish within-group variability from between-group variability, capturing the total variability of the dataset, and are thus well-suited for analysing the corpus of this paper. With a total of 692 titles written by 27 authors in two languages, the textual features that are extracted from the texts must be considered in a hierarchical structure. While linear regression could be used to answer questions such as: 'Do authors construct longer sentences according to their own ageing process?', this would only hold for the analysis of one single author. Chucking the data from all 27 authors together while not accounting for between group variability means that we ignore an author's individual style and linguistic features inherent to different languages.

In addition to the advantages of LMM when working with non-independent data, the model is noted for its robustness when dealing with missing values. Despite the number of available sources reporting on the age of the intended reader as detailed above, this information proved to be unobtainable for several books in the corpus. The sixty titles in question, of which only three written by Dutch or Flemish authors, are all assumed to be targeted towards children or adolescents due their imprint exclusively publishing children's literature. However, the setup of this study is not to look at the dichotomy of literature for adults versus literature for children, but rather investigate a further categorisation of fiction for young readers into narrower age ranges. For this reason, no value is recorded for these sixty books. Common practice dictates

that if linear regression was to be used to analyse the dataset, titles for which the age of the intended reader is unknown would either be removed from the model or their missing value would be estimated based on other entries [29]. However, when working with LMM, these titles are not skipped nor is the age of their intended readers estimated.[1]

5. Results

5.1. Sentence length

In a basic analysis comparing the sentence length of Dutch and English text, a t-test shows language to have a significant effect ($p < 0.0001$). Overall, the average sentence in the English part of the corpus counts 9.051 words, while the mean in Dutch is 8.072 words per sentence. In a LMM fitted with author as a random effect and average sentence length as outcome, the fixed effects of language and the age of the reader, as well as the interaction of both features, are highly significant ($p < 0.0001$). The age of the author proves to be insignificant ($p = 0.056$) and is thus excluded from the model. In the resulting model, the intra-cluster correlation is 0.463 (between author $\sigma = 1.192$; within author $\sigma = 1.283$). When splitting the model by language, the age of the author has a significant effect on the average sentence length of the Dutch part of the corpus ($p = 0.002$).

5.2. Word length

When performing a naive t-test on the average word length, there is a significant difference between Dutch and English texts ($p < 0.0001$). Overall, English words are 4.018 characters long while Dutch words are longer, 4.306 characters. However, the LMM with author as random effect and average word length as outcome does not show a significance for the main effect of language ($p = 0.600$). The age of the author has a significant effect on the average word length ($p < 0.0001$), regardless of language. In terms of average word length, the intra-cluster correlation is 0.35 (between author $\sigma = 0.093$; within author $\sigma = 0.127$). On the other hand, the age of the intended reader and its interaction with language are highly significant ($p < 0.0001$). When splitting the model by language, the effect of the age of the author on the average word length becomes less significant ($p = 0.021$) for Dutch texts while the age of the intended reader becomes slightly less significant ($p = 0.007$) for English texts.

5.3. Lexical diversity

A Welch Two Sample t-test indicates that language has a significant effect on the lexical diversity of texts ($p < 0.0001$) with an average of 0.288 for English and 0.304 for Dutch texts. In a LMM, the age of the author has no significant effect on lexical diversity ($p = 0.061$) and is excluded from the model. Once again, the resulting model shows a highly significant effect of the age of the intended reader ($p < 0.0001$) and a significant interaction between this feature and

[1]The analyses in this paper were conducted using the lme4 package for R. Code and data repository found at: https://zenodo.org/record/7260676.

language (p = 0.011). Language by itself is not a significant effect (p = 0.964). The intra-cluster correlation is 0.500 (between author σ = 0.024; within author σ = 0.023).

5.4. Ratio dialogue vs. narration

Once again, a simple t-test suggests a significant effect of language on the ratio between dialogue and narration. With an average of 0.379, English texts have a higher average value than Dutch texts (0.310). However, none of the fixed effects of language (p = 0.190), age of the intended reader (p = 0.316) and age of the author (p = 0.250) have a significant effect on the ratio between dialogue and narration in a LMM. The correlation between titles of the same author considering this feature is low (0.152).

6. Discussion

A statistical analysis of sentence and word length confirms the hypothesis that the age of the reader has a larger influence on writing style than the individual author or their own age. The correlation between titles of the same author is moderate when considering average sentence length and the age of the author proves to only have a significant effect on the Dutch part of the corpus. Similarly for average word length, the correlation between titles of the same author is low. This supports Jack Grieve's conclusion of his evaluation of authorship attribution techniques; namely that "the value of a single measurement of average word- or sentence-length, [...] appear[s] to be of little use to investigators of authorship" [14]. In contrast to the findings discussed above, the effect of the age of the author on average word length is smaller for Dutch when compared to the English texts. A more significant effect on the average sentence and word length is the age of the intended reader, both isolated and in the interaction with language. This means that the effect of one feature depends on the other. Overall, the LMM estimates that words and sentences lengthen as its readership ages, confirming the hypothesis that these features are closely connected to text complexity and by association to the age of the intended reader. For Dutch texts, 0.02 letters are added to words and sentences become 0.41 words longer when the readership ages with one year. In English, words lengthen with 0.006 letters and only 0.19 words are added to sentences in the same time frame. The effect of the age of the intended readership is thus smaller for the English part of the corpus than for the Dutch.

According to the statistical analyses conducted in this paper, there is a moderate correlation between titles of the same author when considering lexical diversity. Previous research indicates that lexical diversity is correlated with both the age of the author, where complexity decreases with age, and the age of the intended reader, where complexity increases with age. However, in the first case, statistically there seems to be no significant effect between lexical diversity and the age of the author. The hypothesis does prove true, however, for the second case; there is a significant effect of the age of the intended reader as well as of language and the interaction between both features on the lexical diversity of texts in the corpus. When the LMM is split according to language, the model estimates that per year the intended reader of Dutch texts ages, the lexical diversity increases with 0.005941 units. This effect is smaller in the English texts included in the corpus, where the lexical diversity increases with only 0.004358

units. While the difference between these measurements seems small (0.001583), it is statistically significant for an average lexical diversity of 0.3037 for Dutch and 0.2883 for English texts. Thus, lexical diversity of Dutch texts is influenced by the age of the intended reader to a higher degree than English texts.

Statistically there is no significant effect of language, age of the intended reader or age of the author on the ratio between dialogue and narration. Furthermore, there is only a low correlation between measurements of titles of the same author.

7. Conclusion

While the textual features of average sentence length, average word length and lexical diversity are used both in authorship attribution and readability formulas, suggesting that writing style and text complexity are closely connected, a statistical analysis conducted on 692 Dutch and English texts written for children, young adults, and adults suggests that these elements are more related to categories based on the age of the intended reader. None of the features result in a strong correlation between titles of the same author. This supports the fact that the average sentence and word length have often been disputed as reliable for authorship attribution because they can be consciously manipulated by the author to produce a text with a desired level of readability in connection with the age of the intended reader of said text. However, lexical diversity, which is presented by some scholars as an element related to chronological issues, such as the age of the author, also turns out to be linked more closely to the categories determined by the age of the intended reader. The ratio between dialogue and narration, which was included in the analyses as a control feature, shows no correlation with any of the categories studied in this paper. In naive analyses using linear regression models, language has a statistically significant effect on all the textual features included in this study. However, the LMM which takes into account the author as a random effect refutes this conclusion; only the average sentence length is significantly influenced by the difference between Dutch and English texts. Haverals et al. took an important first step in gaining a deeper understanding of the interaction between the writing style and intended reader of a text. This paper built on that by showing that, next to function words, average sentence length, word length and lexical diversity are dependable features for readership attribution. Furthermore, it presented a method that is reliable for further singling out the set of features relevant to determine the age of the intended reader of a text.

Acknowledgments

The author wrote this article as part of the research project Constructing Age for Young Readers. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 804920). The author would like to thank Vanessa Joosen, Mike Kestemont and Wouter Haverals for their support in developing the research that forms the basis of this article, as well as the students who helped with the annotations of the primary texts.

References

[1] C. C. Anderson. "Style in Children's Literature: A Comparison of Passages From Books for Adults and for Children". PhD thesis. Kingston, RI, USA: University of Rhode Island, 1984.

[2] J. Bland. *Children's Literature and Learner Empowerment: Children and Teenagers in English Language Education*. A&C Black, 2013.

[3] I. N. Bozkurt, O. Baghoglu, and E. Uyar. "Authorship Attribution". In: *2007 22nd International Symposium on Computer and Information Sciences*. Ieee, 2007, pp. 1–5.

[4] C. E. Chaski. "Empirical Evaluations of Language-Based Author Identification Techniques". In: *Forensic Linguistics* 8 (2001), pp. 1–65.

[5] R. Clark. "Why All the Counting? Feminist Social Science Research on Children's Literature". In: *Children's Literature in Education* 33 (2002), pp. 285–295. DOI: 10.1023/a:1021276729780.

[6] M. A. Covington. *MATTR User Manual*. Athens, GA: University of Georgia Artificial Intelligence Center, 2007.

[7] M. A. Covington and J. D. McFall. "Cutting the Gordian Knot: The Moving-Average Type-Token Ratio (MATTR)". In: *Journal of Quantitative Linguistics* 17.2 (2010), pp. 94–100. DOI: 10.1080/09296171003643098.

[8] T. Crisp, S. M. Knezek, M. Quinn, G. E. Bingham, K. Girardeau, and F. Starks. "What's on Our Bookshelves? The Diversity of Children's Literature in Early Childhood Classroom Libraries". In: *Journal of Children's Literature* 42.4 (2016), p. 29.

[9] A. B. Diekman and S. K. Murnen. "Learning to be Little Women and Little Men: The Inequitable Gender Equality of Nonsexist Children's Literature". In: *Sex Roles* 50.5 (2004), pp. 373–385. DOI: 10.1023/B:SERS.0000018892.26527.e.

[10] R. S. Forsyth, D. I. Holmes, and E. K. Tse. "Cicero, Sigonio, and Burrows: Investigating the Authenticity of the *Consolatio*". In: *Literary and Linguistic Computing* 14.3 (1999), pp. 375–400.

[11] E. Fry. "Readability versus Leveling". In: *The Reading Teacher* 56.3 (2002), pp. 286–291.

[12] R. Ghesquière. *Het verschijnsel jeugdliteratuur*. Leuven: Acco, 2000.

[13] C. Gregori-Signes and B. Clavel-Arroitia. "Analysing Lexical Density and Lexical Diversity in University Students' Written Discourse". In: *Procedia-Social and Behavioral Sciences* 198 (2015), pp. 546–556.

[14] J. Grieve. "Quantitative Authorship Attribution: An Evaluation of Techniques". In: *Literary and Linguistic Computing* 22.3 (2007), pp. 251–270.

[15] W. Haverals, L. Geybels, and V. Joosen. "A Style for Every Age: A Stylometric Enquiry into Crosswriters for Children, Adolescents and Adults". In: *Language and Literature* 31.1 (2022), pp. 62–84.

[16] J. S. Hillman. "An Analysis of Male and Female Roles in Two Periods of Children's Literature". In: *The Journal of Educational Research* 68.2 (1974), pp. 84–88.

[17] D. I. Holmes. “The Evolution of Stylometry in Humanities Scholarship”. In: *Literary and Linguistic Computing* 13.3 (1998), pp. 111–117.

[18] D. van Hulle and M. Kestemont. “Periodizing Samuel Beckett’s Work: A Stylochronometric Approach”. In: *Style* 50.2 (2016), pp. 172–202.

[19] V. Johansson. “Lexical Diversity and Lexical Density in Speech and Writing: A Developmental Perspective”. In: *Working Papers* 53 (2008), pp. 61–79.

[20] W. Johnson. *Language and Speech Hygiene*. 1939.

[21] P. Juola. *How a Computer Program Helped Show J.K. Rowling Write A Cuckoo’s Calling*. 2013. URL: https://www.scientificamerican.com/article/how-a-computer-program-helped-show-jk-rowling-write-a-cuckoos-calling/.

[22] G. Kjetssa. “And Quiet Flows the Don Through the Computer”. In: *Association for Literary and Linguistic Computing Bulletin* 7 (1978), pp. 248–256.

[23] C. Lennon and H. Burdick. *The Lexile Framework as an Approach for Reading Measurement and Succes*. 2004. URL: https://cdn.lexile.com/m/resources/materials/Lennon%5C%5F%5C%5FBurdick%5C%5F2004.pdf.

[24] V. Levitsky and Y. P. Melnyk. “Sentence Length and Sentence Structure in English Prose”. In: *Glottometrics* 21 (2011), pp. 14–24.

[25] E. F.-L. Y. Ma and R. Loftus. “Ranking-Based Readability Assessment for Early Primary Children’s Literature”. In: *Human Language Technologies*. Montréal: Association for Computational Linguistics, 2012, pp. 548–552.

[26] D. Mannion and P. Dixon. “Sentence-Length and Authorship Attribution: The Case of Oliver Goldsmith”. In: *Literary and Linguistic Computing* 19.4 (2004), pp. 497–508. DOI: 10.1093/llc/19.4.497.

[27] T. C. Mendenhall. “The Characteristic Curves of Composition”. In: *Science* 11 (1887), pp. 237–249.

[28] B. J. Meyer. “Identification of the Structure of prose and Its Implications for the Study of Reading and Memory”. In: *Journal of Reading Behavior* 7.1 (1975), pp. 7–47.

[29] G. Molenberghs, L. Bijnens, and D. Shaw. “Linear Mixed Models and Missing Data”. In: *Linear Mixed Models in Practice*. Vol. 126. Lecture Notes in Statistics. New York, NY: Springer-Verlag, 1997, pp. 191–274. DOI: 10.1007/978-1-4612-2294-1_5.

[30] R. G. Potter. “Statistical Analysis of Literature: A Retrospective on Computers and the Humanities, 1966-1990”. In: *Computers and the Humanities* 25.6 (1991), pp. 401–429.

[31] D. Tallentire. “Confirming Intuitions About Style, Using Concordances”. In: *The Computer in Literary and Linguistic Studies*. Ed. by A. Jones and R. Churchhouse. Cardiff: University of Wales Press, 1976, pp. 309–338.

[32] F. Wanner, J. Fuchs, D. Oelke, and D. A. Keim. “Are my Children Old Enough to Read these Books? Age Suitability Analysis”. In: *Polibits* 43 (2011), pp. 93–100.

[33] G. U. Yule. “On Sentence-Length as a Statistical Characteristic of Style in Prose, with Application to Two Cases of Disputed Authorship”. In: *Biometrika* 30 (1938), pp. 363–390.